CROSS − VALIDATION :—

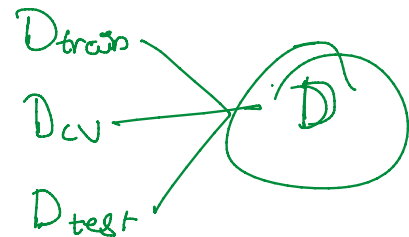We typically split data into D_{train} & D_{test}.

We train model on D_{train} & we get the accuracy on D_{test}.

But we can't generalize this result on future data. Because to evaluate / to do hyper parameter tuning we have used D_{test} here.

So we can't generalize the result on future unseen data.

So we need to have some other idea.

We can split our data into

D_{train}
D_{cv} — D
D_{test}

D_{cv} → Cross validation data.

D_{train} ——→ Train the model.

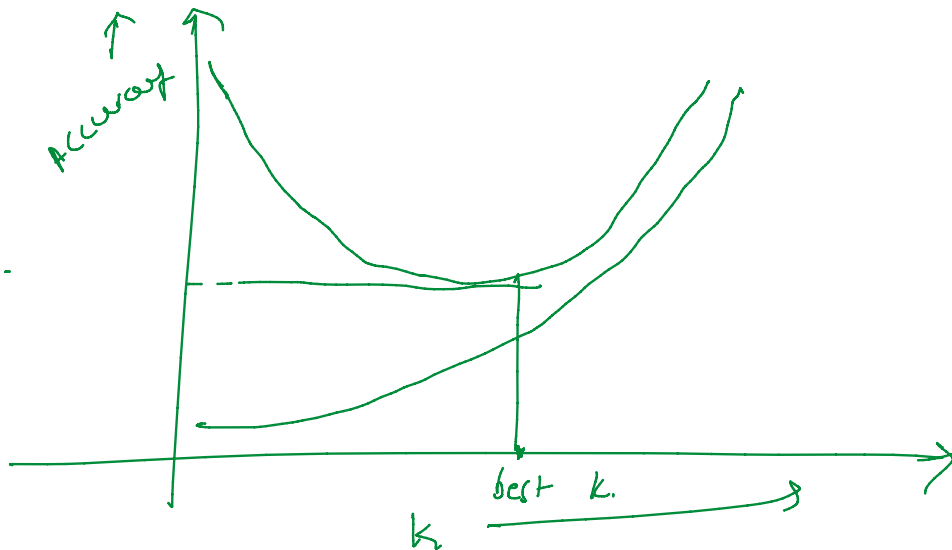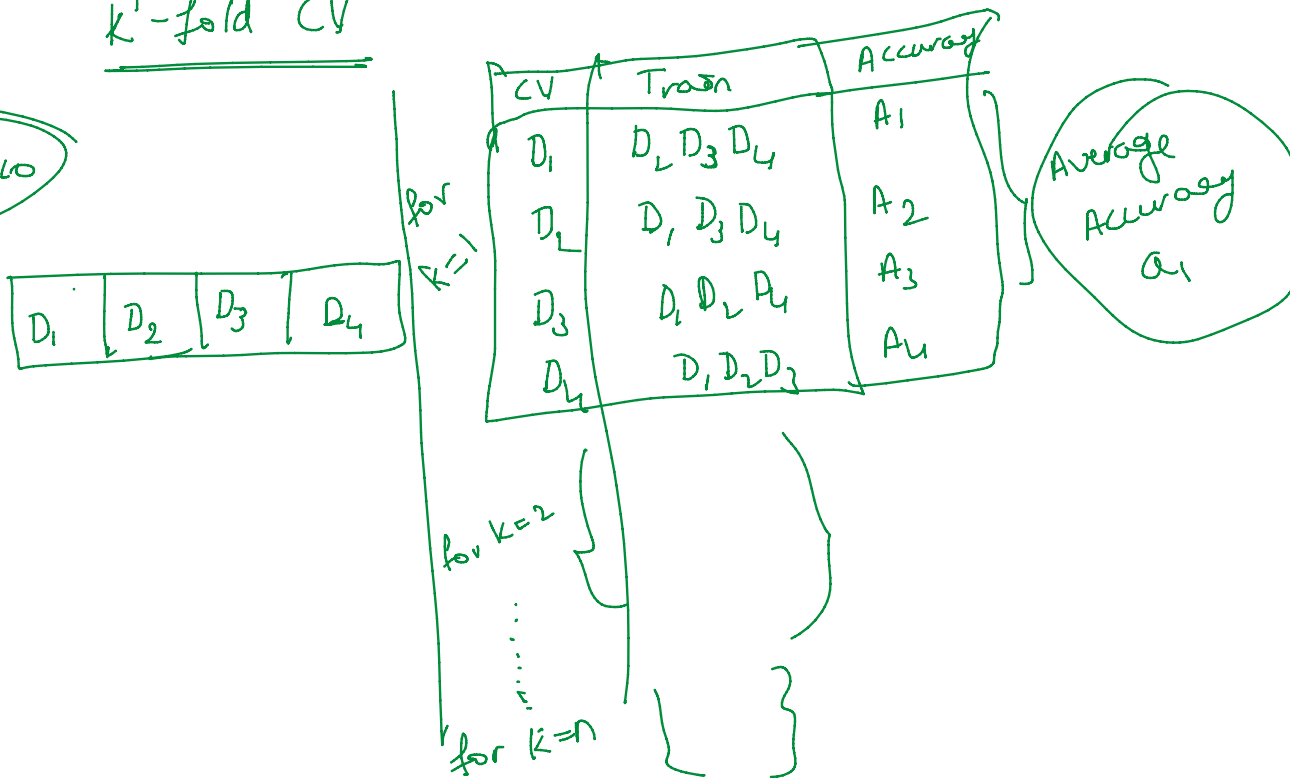D_{cv} ——→ Hyper parameter tuning

D_{test} ——→ Used as unseen data for generalization.

By splitting train Data into train & CV, we loose ___ for training. Hence for better performance

By splitting some data for training. Hence for better performance we use, **k'-fold CV**

typically $k' = 10$

| D_1 | D_2 | D_3 | D_4 |

for $K=1$

CV	Train	Accuracy
D_1	$D_2\ D_3\ D_4$	A_1
D_2	$D_1\ D_3\ D_4$	A_2
D_3	$D_1\ D_2\ D_4$	A_3
D_4	$D_1\ D_2\ D_3$	A_4

Average Accuracy a_1

for $K=2$

. . . . . .

for $K=n$

Accuracy ↑ / k →   best k.

So, It is best practice to perform cross validation to find best hyperparameter.

best hyperparameter.